

Compulsory exercise 1: Group 28

TMA4250 Spatial Statistics V2022

Nora Røhnebæk Aasen, Elias Klakken Angelsen

18 februar, 2022

Problem 1: GRFs - model characteristics

a)

The correlation function ρ is supposed to be a positive semi-definite function. This means that we require

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j \rho(s_i, s_j) \geq 0, \quad \forall a_1, \dots, a_m \in \mathbb{R}, \forall s_1, \dots, s_m \in \mathcal{D}$$

This is a requirement in order to ensure well-defined (i.e. positive) variance, since

$$\text{Var} \left(\sum_{i=1}^m b_i X(s_i) \right) = \sum_{i=1}^m \sum_{j=1}^m b_i b_j C(s_i, s_j) = \sum_{i=1}^m \sum_{j=1}^m b_i b_j \sigma^2 \rho(s_i, s_j) = \sum_{i=1}^m \sum_{j=1}^m a_i a_j \rho(s_i, s_j),$$

where C is the covariance function, $\sigma^2 = C(0)$ and $a_i = b_i \sigma$ and $a_j = b_j \sigma$.

Below we plot the correlation function $\rho(h) = C(h)/C(0)$ and the semi-variogram function $\gamma(h) = \frac{1}{2}[\text{Var}(X(h) - X(0))]$ for 4 different covariance functions, and for two different choices of σ^2 .

The two covariance models are Matérn (note that this model differs from the one used in `cov.spatial` in R, since what they call ϕ is $\frac{a}{\sqrt{8\nu}}$ for us):

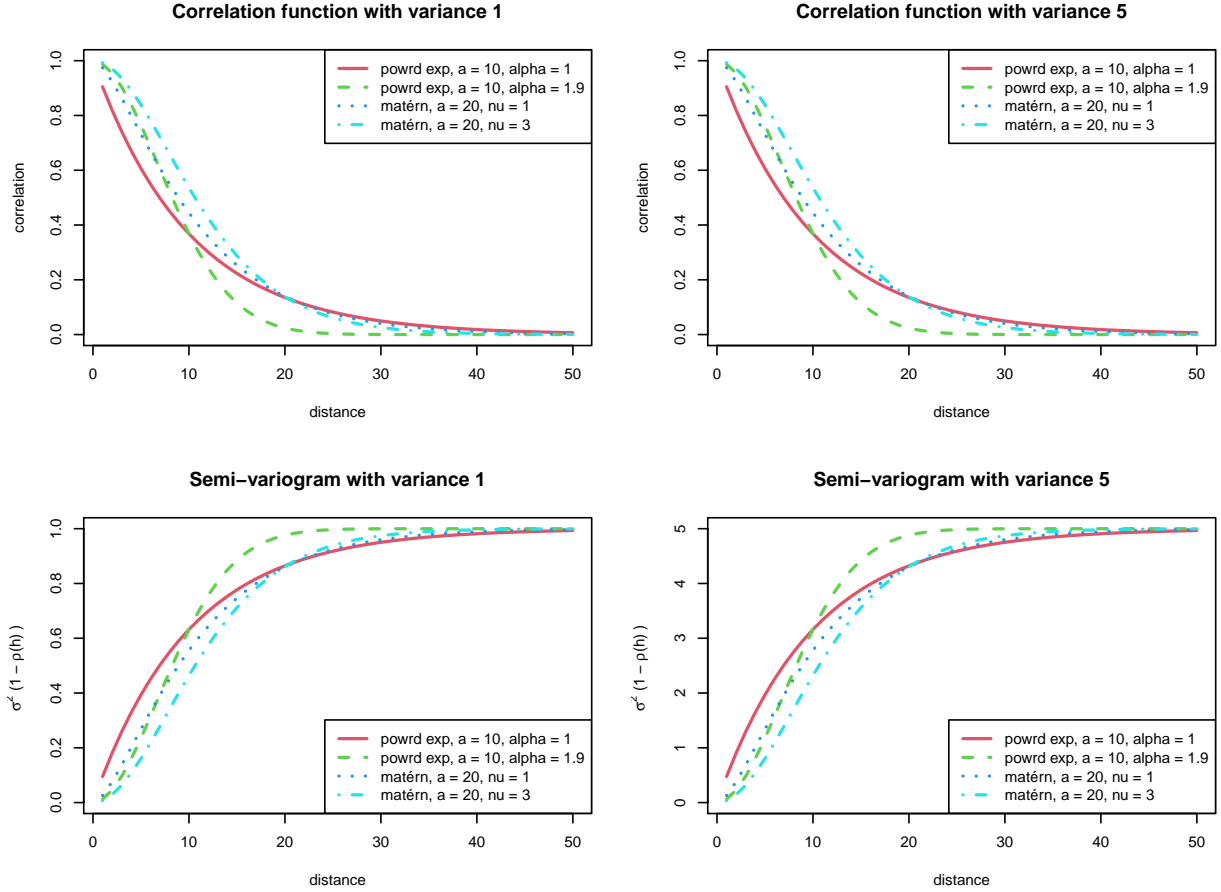
$$C(h; a, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{8\nu} h}{a} \right)^\nu K_\nu \left(\frac{\sqrt{8\nu} h}{a} \right)$$

and powered exponential:

$$C(h; a, \alpha) = \exp \left\{ - \left(\frac{h}{a} \right)^\alpha \right\}, \quad 0 < \alpha \leq 2$$

These can be plotted in one dimension since they are stationary in this case, and, additionally, isotropic since $\mathcal{D} \subset \mathbb{R}$.

We plot the correlation functions and the variograms for our models.



The correlation plots are equal, as the varying factor σ^2 doesn't affect the correlation. The semi-variogram depends on the covariance function through $\gamma(h) = [C(0) - C(h)]$, and since we have $C(0) = \sigma^2$ and $\rho(h) = C(h)/C(0)$, we can show

$$\gamma(h) = [C(0) - C(h)] = [\sigma^2 - \sigma^2 \rho(h)] = \sigma^2[1 - \rho(h)].$$

This relationship is visible through the plots. Furthermore the range of the covariance function and the sill can be read from the semi-variogram plots. The range is the distance at which dependence is negligible, which is 10 and 20 in our case respectively. We see that the Matérn covariances has a longer range than the powered exponential. The sill is the marginal variance, i.e. σ^2 , and can be found around the point where the semi-variogram decays. In our case, this is 1 and 5 as expected.

The behavior around 0 indicates the differentiation of the covariance function and whether it is quadratic mean continuous or not. From the graph we would assume the powered exponential with range 10 and $\alpha = 1$ is not continuous, while the others are. We need $2k$ times differentiability in 0 for the covariance function to be k times quadratic mean differentiable. By looking we could guess that Matérn with range 20 and $\nu = 1$ is not differentiable, whereas the two last ones might be at least one time quadratic mean differentiable.

For the variogram function $2\gamma(h)$, the relationship to the covariance function is

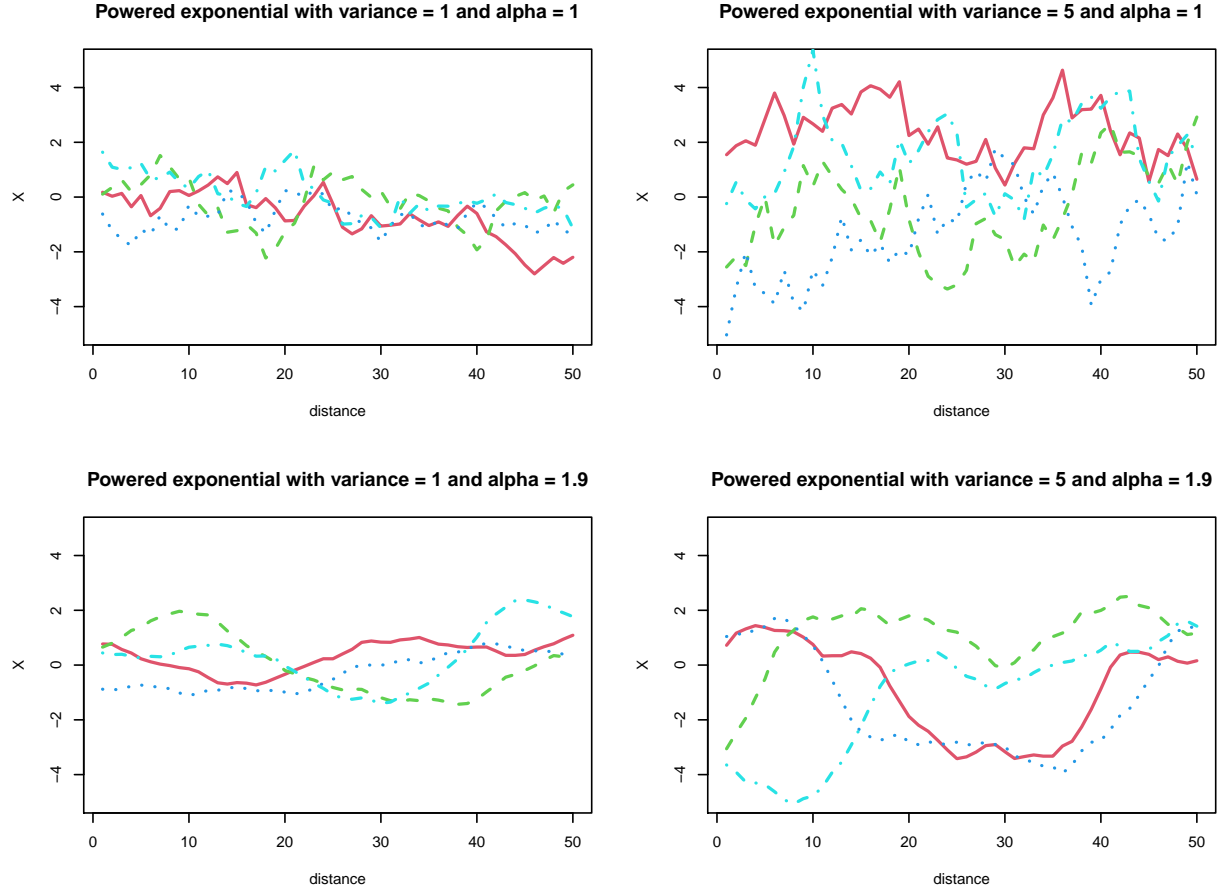
$$2\gamma(h) = 2[C(0) - C(h)] = 2\sigma^2[1 - \rho(h)].$$

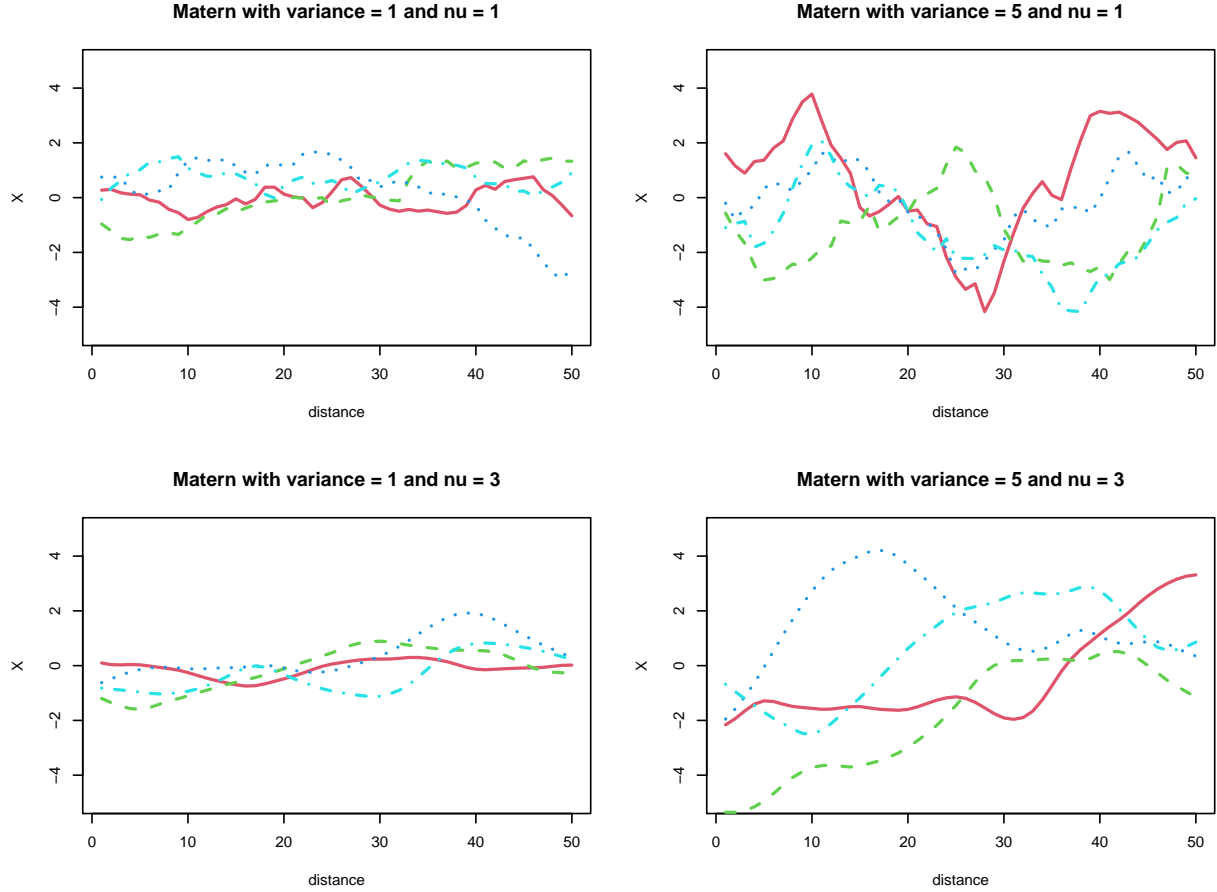
b)

Since X is a Gaussian random field, we know that its distribution is completely determined by its mean and covariance matrix. Furthermore, the vector $\mathbf{X} = (X(s_1), \dots, X(s_{50}))$ is also normal and has a mul-

tivariate normal distribution. Therefore, $\mathbf{X}(s) \sim \mathcal{N}_{50}(\mu, \Sigma)$, where $\mu = E[\mathbf{X}(s)] = 0$ in our case, and $\Sigma_{ij} = Cov(X(s_i), X(s_j)) = \sigma^2 \rho(\|s_i - s_j\|)$ for $s \in \mathcal{D}$.

We simulate four realizations from each model.





As we can see, the plots depend both on α (in the powered exponential) or ν (in the Matern model) and σ^2 . Increased marginal variance leads to higher variance in the realizations, as expected.

In the model with the Matern covariance function, changing ν often results in changing the (mean squared) differentiability of the process. One can show that the process will be $\lceil \nu \rceil - 1$ times differentiable. Hence, changing ν from 1 to 3 takes the differentiability from zero times differentiable to two times differentiable, as we can anticipate from the plots.

If we would let ν go to infinity, it would approach a Gaussian covariance function (powered exponential with $\alpha = 2$) and for $\nu = 0.5$, we would get the exponential covariance function ($\alpha = 1$).

Hence, α controls the degree of differentiability of the process with powered exponential covariance function, as we can see from the plots. As $\alpha \rightarrow 2$, it becomes smooth, while for $\alpha = 1$, it would not even be differentiable.

We also had the parameter $a = 10$ for the powered exponential and $a = 20$ for the Matern. This parameter controls how fast the correlation decays with distance. As we have a higher a for the Matern, we could expect that it fluctuates less over distance, which is also what we observe in the realizations. The powered exponential models (where $a = 10$) go more “up and down”, while the Matern models (where $a = 20$) “stays flat” more often.

c)

We set $s_1 = 10, s_2 = 25$ and $s_3 = 30$. At these points, the observations are given by $Y_i = X(s_i) + \varepsilon_i$ for $i = 1, 2, 3$, where $\varepsilon_1, \varepsilon_2, \varepsilon_3 \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_N^2)$. We also have $X(s_i) \sim \mathcal{N}(0, \sigma^2)$ for each i .

Hence, each Y_i is distributed as $Y_i \sim \mathcal{N}(0, \sigma^2 + \sigma_N^2)$ as the errors ε_i are independent of the $X(s_i)$'s.

This implies that the vector $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$ is distributed as a multivariate normal distribution, $\mathbf{Y} \sim \mathcal{N}_3(0, \Sigma)$, where the covariance matrix Σ is given by

$$\begin{aligned}\Sigma_{ij} &= \text{Cov}[Y_i, Y_j] = \text{Cov}[X(s_i) + \varepsilon_i, X(s_j) + \varepsilon_j] \\ &= \text{Cov}[X(s_i), X(s_j)] + \text{Cov}[X(s_i), \varepsilon_j] + \text{Cov}[\varepsilon_i, X(s_j)] + \text{Cov}[\varepsilon_i, \varepsilon_j] \\ &= \begin{cases} \sigma^2 + \sigma_N^2 & i = j \\ \sigma^2 \rho(\|s_i - s_j\|) & i \neq j. \end{cases}\end{aligned}$$

d)

We consider the Matern covariance model with $\sigma^2 = 5$ and $\nu = 3$ and choose the first realization.

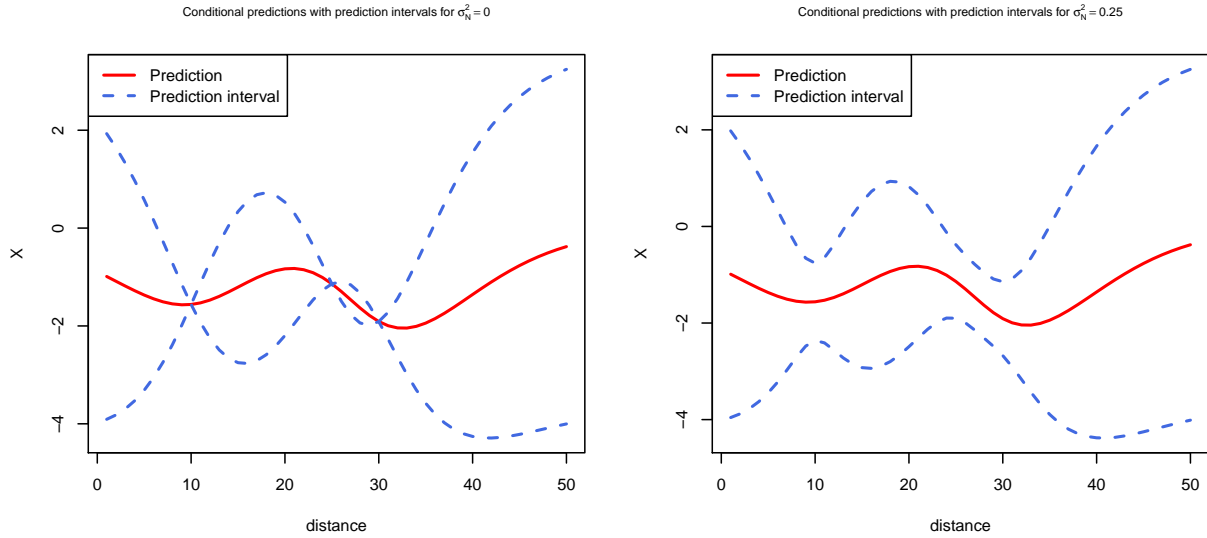
We consider first the vector $[\mathbf{X}, \mathbf{Y}]^T \sim \mathcal{N}([\mu_0, \mu_1]^T, \Sigma)$, where $[\mu_0, \mu_1]^T = [0, 0]^T$ and $\Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}$. Here Σ_{00} and Σ_{11} are the covariance matrices of \mathbf{X} and \mathbf{Y} respectively and $\Sigma_{01} = \Sigma_{10}^T = [\text{Cov}(X_i, Y_j)]_{ij}$ for $i = 1, \dots, 50$ and $j = 1, 2, 3$.

We can calculate the covariance as follows,

$$\text{Cov}(X_i, Y_j) = \text{Cov}(X_i, X_j + \varepsilon_j) = \text{Cov}(X_i, X_j) + \text{Cov}(X_i, \varepsilon_j) = \text{Cov}(X_i, X_j).$$

Hence, $\Sigma_{01} = \Sigma_{10}^T = [\text{Cov}(X_i, X_j)]_{ij}$ for $i = 1, \dots, 50$ and $j = 10, 25, 30$.

The conditional distribution can be shown to be given by $\mathbf{X}|\mathbf{Y}=y \sim \mathcal{N}(\mu_0 + \Sigma_{01}\Sigma_{11}^{-1}(y - \mu_1), \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10})$.



To make the prediction intervals, we used the known mean and variance and the properties of the normal distr. i.e. $\frac{X_{pred} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ to calculate the upper and lower bounds of the prediction intervals, i.e.

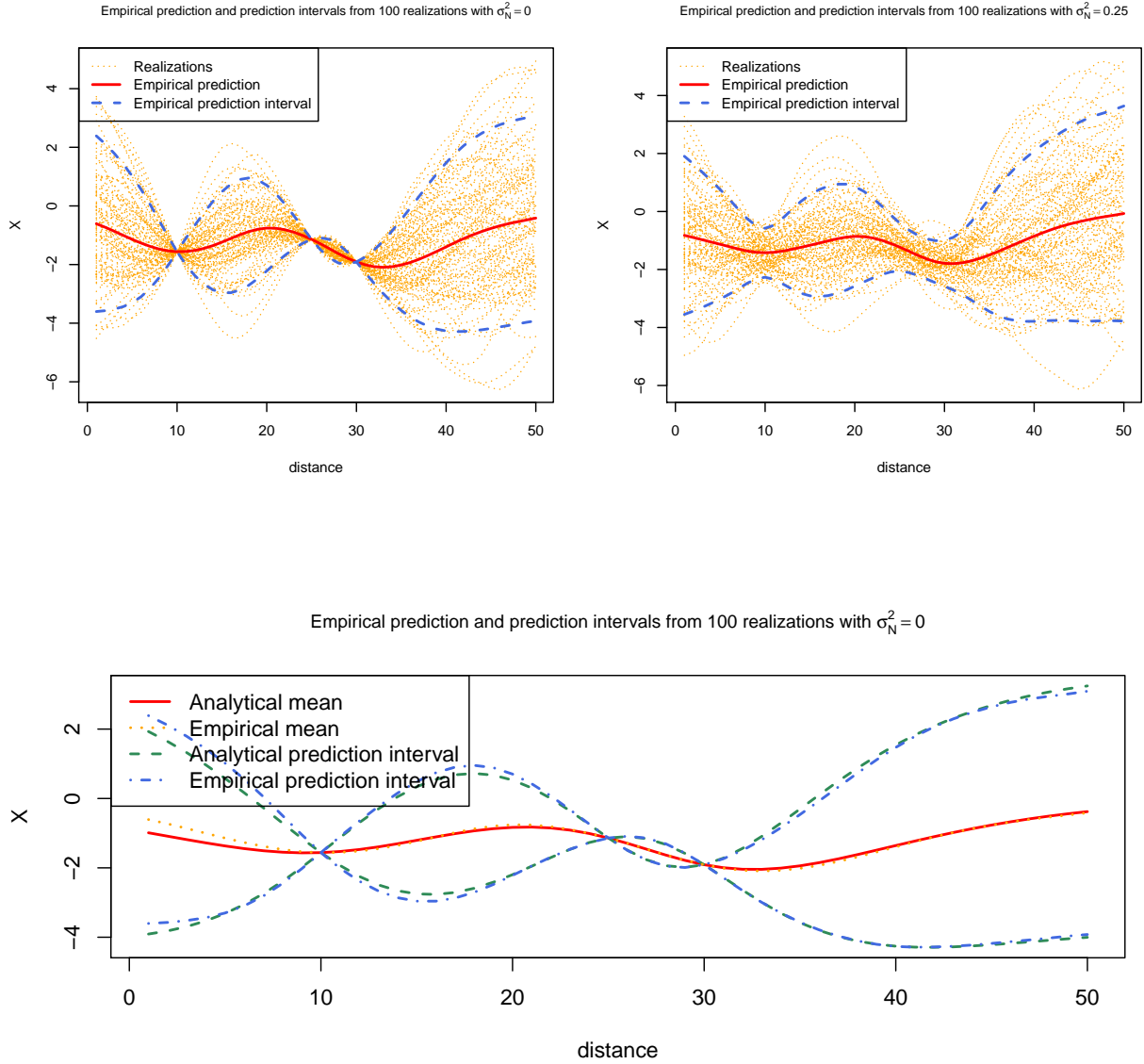
$$P(-z_{0.05} \leq \frac{X_{pred} - \mu}{\sigma} \leq z_{0.05}) = 90\% \implies P(X_{pred} \in [\mu - z_{0.05}\sigma, \mu + z_{0.05}\sigma]) = 90\%.$$

We found σ^2 by taking the diagonal of the covariance matrix.

From both plots it is clear that the observation decrease the variance. In the left figure $\sigma_N^2 = 0$, hence there is no uncertainty in the points 10, 25 and 30. The variance is lower at this points in the right plot as well, but since there is some uncertainty in the observations the prediction interval is wider here. At $s = 0$ and $s = 50$ the prediction intervals are the widest, as the observations are too far away to give any accurate indication of the values at the ends.

e)

We now compute 100 realizations from the conditional model $\mathbf{X}|\mathbf{Y}=y \sim \mathcal{N}(\mu_0 + \Sigma_{01}\Sigma_{11}^{-1}(y - \mu_1), \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10})$, and use these to estimate a prediction of X with prediction intervals.



To make the prediction intervals, we had to calculate the empirical mean and variance. And then we used the property that $\frac{X_{pred} - \bar{X}}{\sqrt{s^2(1+1/n)}} \sim t_{n-1}$ to calculate the upper and lower bounds of the prediction intervals as

follows

$$P(-t_{99,0.05} \leq \frac{X_{pred} - \mu}{s} \leq t_{99,0.05}) = 90\% \implies P(X_{pred} \in [\mu - t_{99,0.05}s, \mu + t_{99,0.05}s]) = 90\%,$$

where s is the estimated standard deviation.

The empirical prediction interval should contain approximately 90% of our observations, which seems true. Furthermore, we again see the trend of lower variance close to the observed points, and higher variance further away. From the last plot we can compare the empirical and analytical results, and we see that they are pretty consistent. This is not too surprising, as 100 realizations is a quite good sample size. In general the empirical prediction interval is wider, which is what we would expect. However, from the last observation and out, i.e. for $s \in [30, 50]$, the empirical prediction interval is actually smaller. This might be a result of “lucky” realizations, where less than expected are outside the analytical prediction interval, hence underestimating the variance.

f)

We now want to estimate the expected size of the area where $X > 2$ in two ways. First, let $A = \sum_{s \in \tilde{\mathcal{D}}} \mathbb{I}(X(s) > 2)(X(s) - 2)$ be a function of $X(s)$. One way to find the expected size of this area is to calculate $E[A(X(s))] = \hat{A}$.

Another method is to calculate $\tilde{A} = A(\hat{X})$, where \hat{X} is the simple kriging predictor of X . The simple Kriging predictor at a location s_0 is given by $\hat{X}_0 = c^T \Sigma^{-1} X$, where $X = (X(s_1), \dots, X(s_n))$, $c = \text{Cov}(X, X(s_0))$ and $\Sigma = \text{Cov}(X, X)$.

```
##      A      A Var(A)
## [1,] 0 2.213 20.052
```

We see that \tilde{A} predicts 0 whereas \hat{A} predicts more than 1. This is likely due to the fact that \hat{A} has multiple realizations that pass 2 due to high variance, and this allows for an estimate $\hat{A} > 0$. When using simple kriging we have already found the expected value of X in each point and in no point is our GRF expected to pass 2. The high variance of \hat{A} shows that both estimates agree that it is reasonable for X to never pass 2, i.e. that the area is 0. That $\tilde{A} \leq \hat{A}$ will always be true, and can be shown as follows.

Notice that A is a convex function. Additionally, since X is a GRF we know that the simple kriging predictor of X is also the expectation of X in that point, i.e. $\hat{X} = E[X]$.

Using Jensen’s inequality, which says that

$$g(E[X]) \leq E[g(X)],$$

it follows immediately that $\tilde{A} \leq \hat{A}$. To see this, note that

$$\tilde{A} = A(\hat{X}) = A(E[X]) \leq E[A(X)] = \hat{A},$$

by Jensen’s inequality.

g)

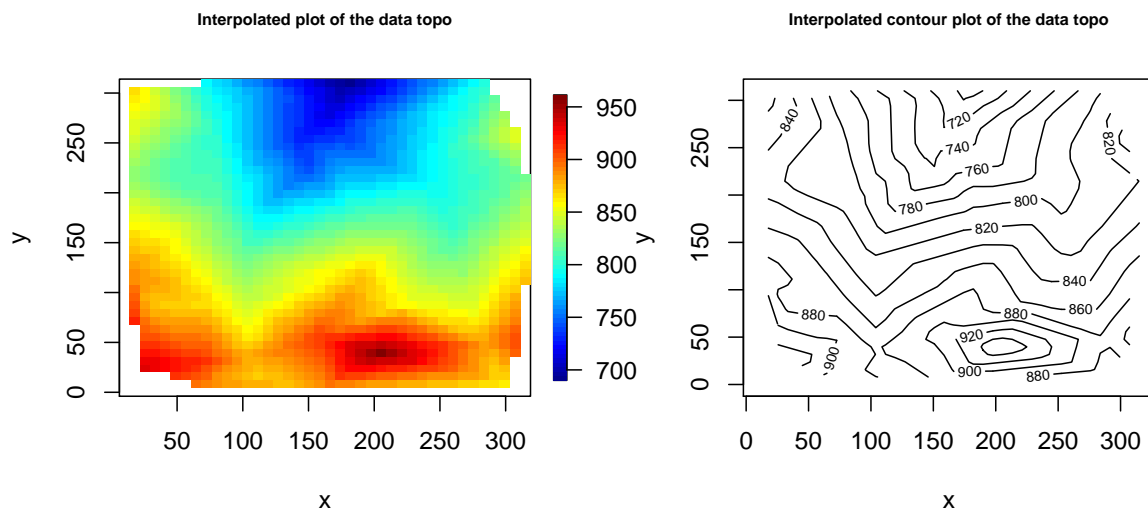
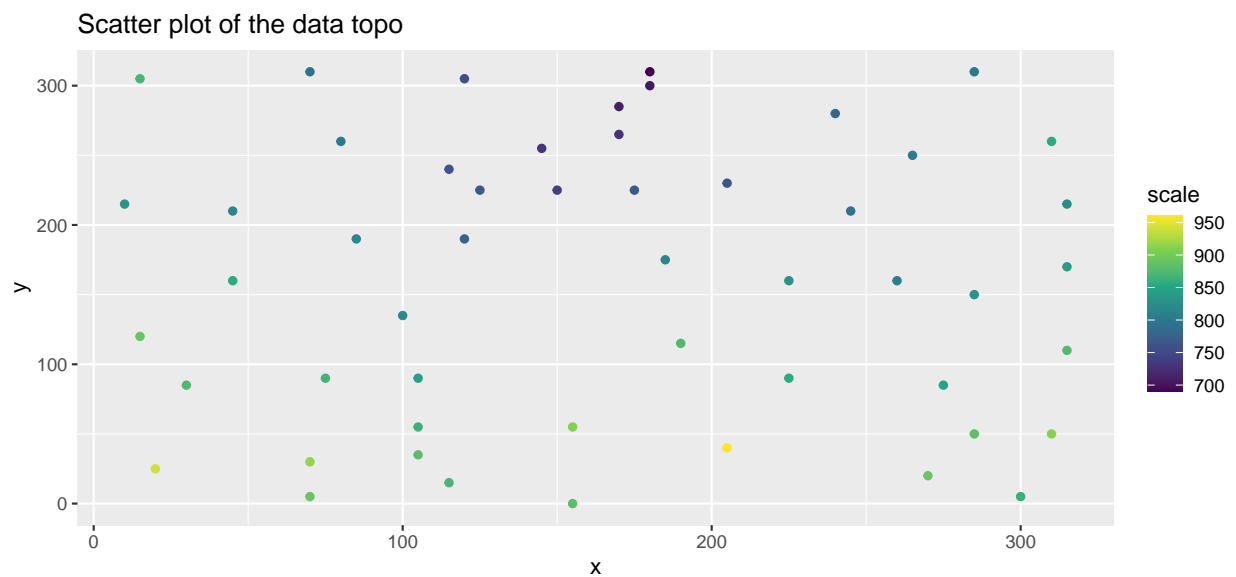
In task one we have seen that the parameters ν for Matérn and α for powered exponential determines the smoothness of our covariance function, whereas the marginal variance σ^2 and range a determines the variance of our function. A high marginal variance σ^2 can be “held in check” if the range too is high. Thus, the potential variance is high but a realization of the RF will change slowly since the range restricts it from behaving very unpredictable.

Furthermore, we noticed that observations can significantly reduce variance of our predictions. However, when the range is small and the variance is high, if we only observe a few data points we should not be overconfident in estimates far from our observations.

Lastly, we saw that if we want to calculate something from our observations we have to be careful of whether we choose to find a prediction based on our observations and then calculate what we are interested in, or if we first choose to calculate what we want based on all observations and then predict our calculation from this. These two methods will not yield the same answer.

Problem 2

a)



From the plots we would not assume that this is a stationary GRF since there is no clear consistency in the covariance over certain distances over the different areas in the plot. For instance, the upper region of the plot seem to have more dependency than the lower region. In the scatter plot as well we see that the yellow point above 200 on the x-axis seem very odd, and not displaying the strong dependence to its neighbouring points as seen for the dots in the upper region.

b)

We now assume that the GRF X on the domain $\mathcal{D} \in [0, 350] \times [0, 350] \subset \mathbb{R}^2$ is modeled by

$$E[X(s)] = g(s)^T \beta, \quad s \in \mathcal{D} \quad Var[X(s)] = \sigma^2 = 50^2, \quad s \in \mathcal{D} \quad Corr[X(s), X(s')] = \rho(\|s-s'\|) \stackrel{h=\|s-s'\|}{=} e^{-(0.01h)^{1.5}}, \quad h \in [0, \infty),$$

We recognize the correlation function as a powered exponential.

If we want to predict an unobserved value of X at the point s_0 , we could use universal kriging. Universal kriging finds the BLUP, i.e. the best linear unbiased estimate \hat{X}_0 of $X(s_0) = X_0$. This means we are looking for \hat{X}_0 satisfying the following:

- **Best:** The mean square error $E[(X_0 - \hat{X}_0)^2]$ is minimized.
- **Unbiased:** $E[\hat{X}_0] = E[X_0]$
- **Linear:** $\hat{X}_0 = \mathbf{a}^T \mathbf{X}$, i.e. it is a linear combination of our observations.

From this we see that we get a constraint on β from

$$\begin{aligned} E[\hat{X}_0] &= E[X_0] \\ E[\mathbf{a}^T \mathbf{X}] &= E[X_0] \\ \mathbf{a}^T Z^T \beta &= Z_0^T \beta \\ \mathbf{a}^T Z^T \beta - Z_0^T \beta &= 0 \\ (\mathbf{a}^T Z^T - Z_0^T) \beta &= 0 \end{aligned}$$

where the mean function $m(s) = \sum_{k=1}^p f_k(s) \beta_k = Z_s^T \beta$ and $Z = (Z_1, \dots, Z_n)$.

From this we can minimize the MSE using Lagrange multipliers to incorporate the constraint on beta by solving

$$\min_{\beta} (E[(Z_0^T \beta - \mathbf{a}^T Z^T \beta)^2] - \lambda((\mathbf{a}^T Z^T - Z_0^T) \beta))$$

The prediction variance σ_{uk}^2 is given by

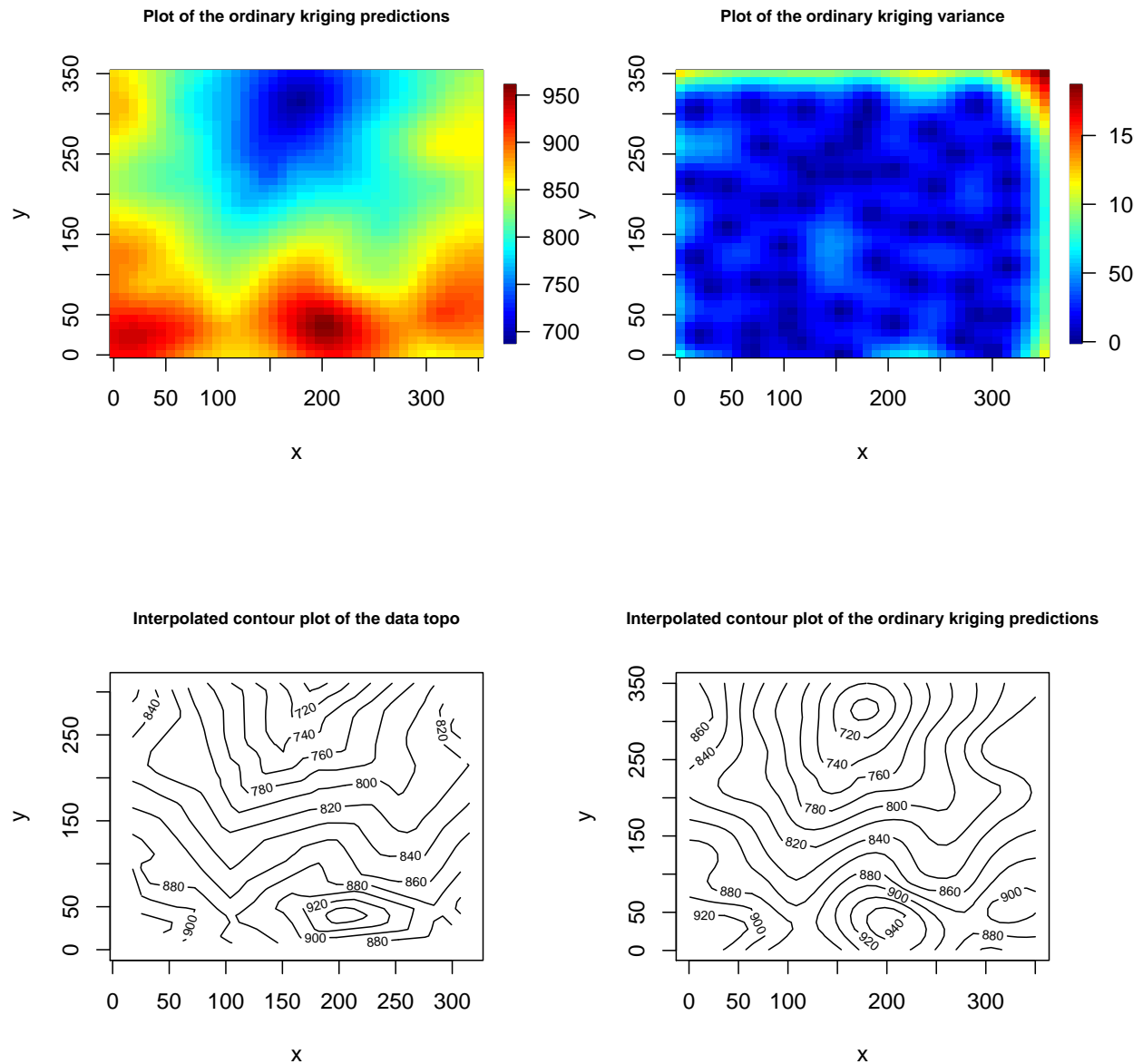
$$\begin{aligned} \sigma_{uk}^2 &= MSE(\hat{\beta}) \\ &= E[(Z_0^T \hat{\beta} - \mathbf{a}^T Z^T \hat{\beta})^2] \\ &= \hat{\beta}^T Var(Z_0) \hat{\beta} + (\mathbf{a} \hat{\beta})^T Var(Z) \mathbf{a} \hat{\beta} - 2Cov(Z_0^T \hat{\beta}, \mathbf{a}^T Z^T \hat{\beta}) \\ &= \hat{\beta}^T Var(Z_0) \hat{\beta} + (\mathbf{a} \hat{\beta})^T Var(Z) \mathbf{a} \hat{\beta} - 2\hat{\beta}^T Cov(Z_0, Z) \mathbf{a} \hat{\beta} \end{aligned}$$

We would expect that the value of σ^2 change for different parametrizations of the expectation function, as less fluctuations of the mean function (i.e. constant or 1st degree linearity) will potentially require more variance in order to explain all the observations, whereas higher degree mean functions may be biased against the observations and therefore explain most of the data without too much variance.

c)

We now assume $E[\mathbf{X}(s)] = \beta_1$, and we want to predict using ordinary Kriging.

```
## krige.conv: model with constant mean
## krige.conv: Kriging performed using global neighbourhood
```



We see that there in general is low variance for our predictors. It's quite impressive that the estimate is this certain from merely 52 observations. We also see that compared with the interpolated contour plot to the left, the interpolation of the ordinary kriging predictions appear more stationary. This is due to the fact that it is an assumption of the model.

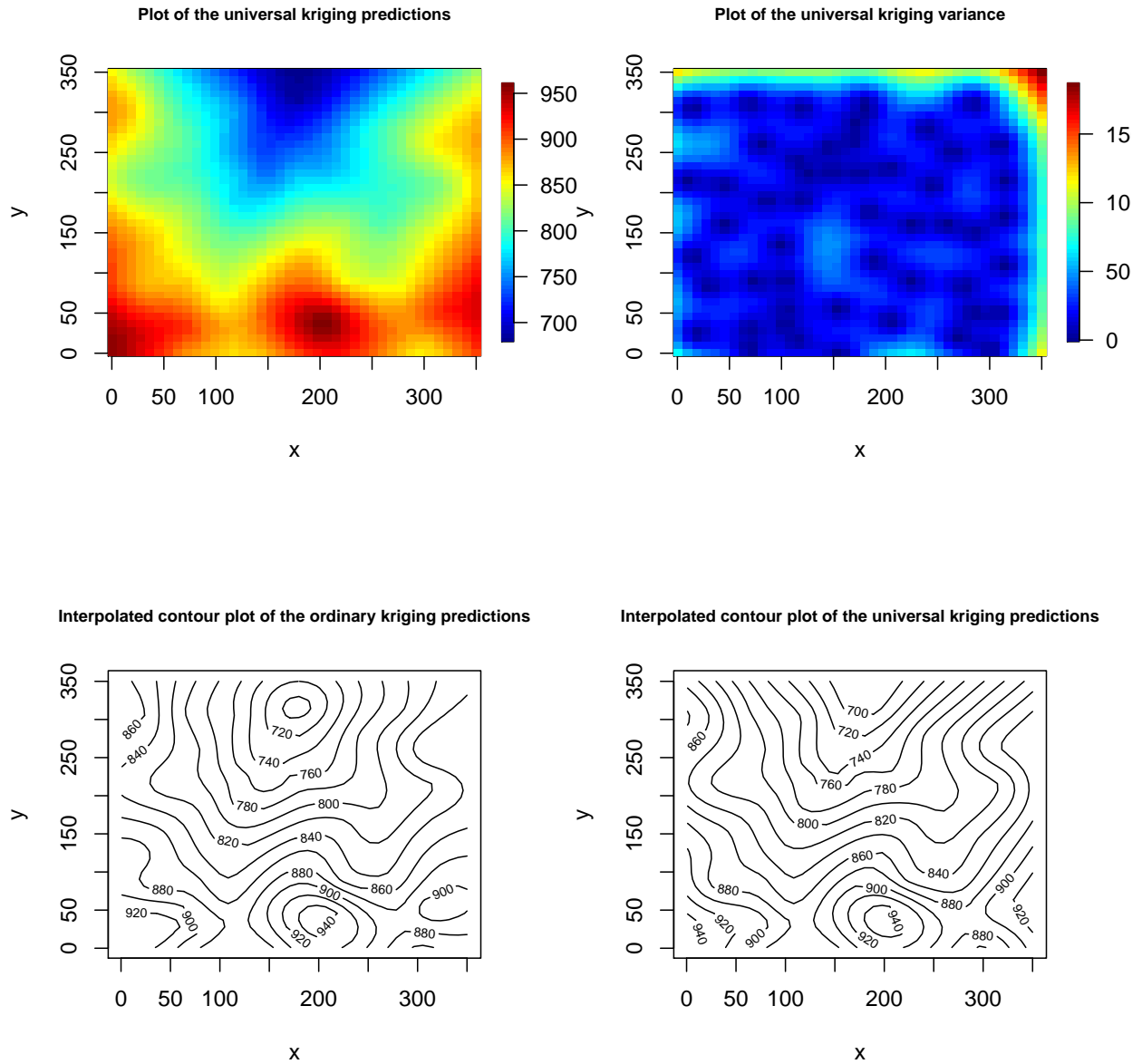
d)

We now let $g(s) = (1, s_1, s_2, s_1 s_2, s_1^2, s_2^2)$. The expected value of $\mathbf{X}(s)$ will then be

$$E[\mathbf{X}(s)] = m(s) = \beta_1 + \beta_2 s_1 + \beta_3 s_2 + \beta_4 s_1 s_2 + \beta_5 s_1^2 + \beta_6 s_2^2$$

krige.conv: model with mean given by a 2nd order polynomial on the coordinates

krige.conv: Kriging performed using global neighbourhood



Again we see that there is quite low variance in our result. An interesting thing to note when we compare universal kriging predictions with the ordinary kriging predictions is that in areas where we have a high value, ordinary kriging does not predict higher values and it also predicts circular height curves around each

high value. This is due to the assumption of constant mean. The universal kriging predictions, on the other hand, has less “stationary-looking” predictions, because the polynomial in the mean allows for more fluctuations. We see it in particular in the middle upper region, and in the bottom left and right corner.

e)

We now consider $s_0 = (100, 100)^T$. Our ordinary kriging estimate at this point is 838.6781414. We can use this prediction along with the prediction variance from our ordinary kriging to compute

$$P(X(s_0) > 850) = P\left(\frac{X(s_0) - \hat{X}_0}{\sigma_{ok}} > \frac{850 - \hat{X}_0}{\sigma_{ok}}\right),$$

which we can solve using the quantiles of a student-t distribution.

From the calculations we see that $P(X(s_0) > 850) = 13.25\%$.

We now want to find the elevation level at which we are 90% sure that the true value $X(s_0)$ is below, i.e. we want to solve $P(X(s_0) < y) = 0.9$ for y . We can rewrite this as $y = t_{51,0.1}\sqrt{\sigma_{ok}} + \hat{X}_0$.

From the calculations we see that 851.72 is the elevation level at which the true value $X(s_0)$ is 90% certainly below.

We see that the upper bound is very close to the predicted value in this point, again indicating quite low variance.

f)

In this task we used a dataset of 52 observations. Both the ordinary kriging and the universal kriging predictors had low variance, but for this problem we would assume that universal kriging is a better method. This is due to the fact that our data didn't seem certainly stationary, and the polynomial in our mean cut some slack on this restriction. We were surprised by how precise the predictions are from merely 52 observations on a grid of size $[1, 315] \times [1, 315]$.

Problem 3 - Parameter estimation

a)

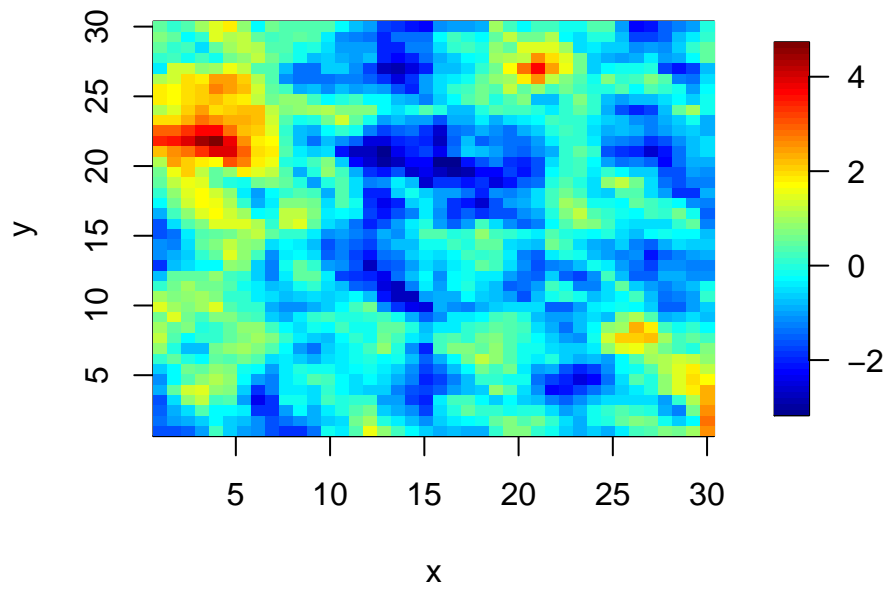
We now consider a stationary GRF $\{X(s)|s \in \mathcal{D}\}$, where $\mathcal{D} = [1, 30]^2$, with

$$E[X(s)] = \mu = 0, Var[X(s)] = \sigma^2 = 2, Corr[X(s), X(s')] = \exp\left(\frac{-\|s - s'\|}{a}\right) \stackrel{h=\|s-s'\|}{=} \exp\left(\frac{-h}{3}\right)$$

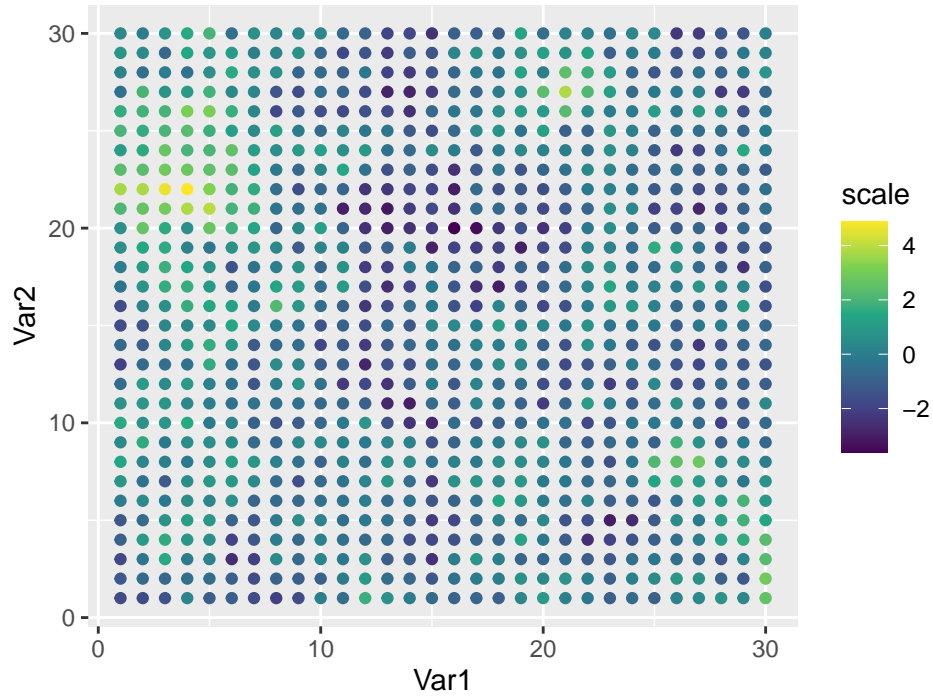
We recognize the correlation function as a powered exponential with $\alpha = 1$ and spatial scale $a = 3$ and find a realization of the GRF.

```
## grf: simulation on a set of locations provided by the user
## grf: process with 1 covariance structure(s)
## grf: nugget effect is: tausq= 0
## grf: covariance model 1 is: powered.exponential(sigmasq=2, phi=3, kappa = 1)
## grf: simulation using the function GaussRF from package RandomFields
## grf: End of simulation procedure. Number of realizations: 1
```

One realization of the GRF X



Scatter plot of the data topo



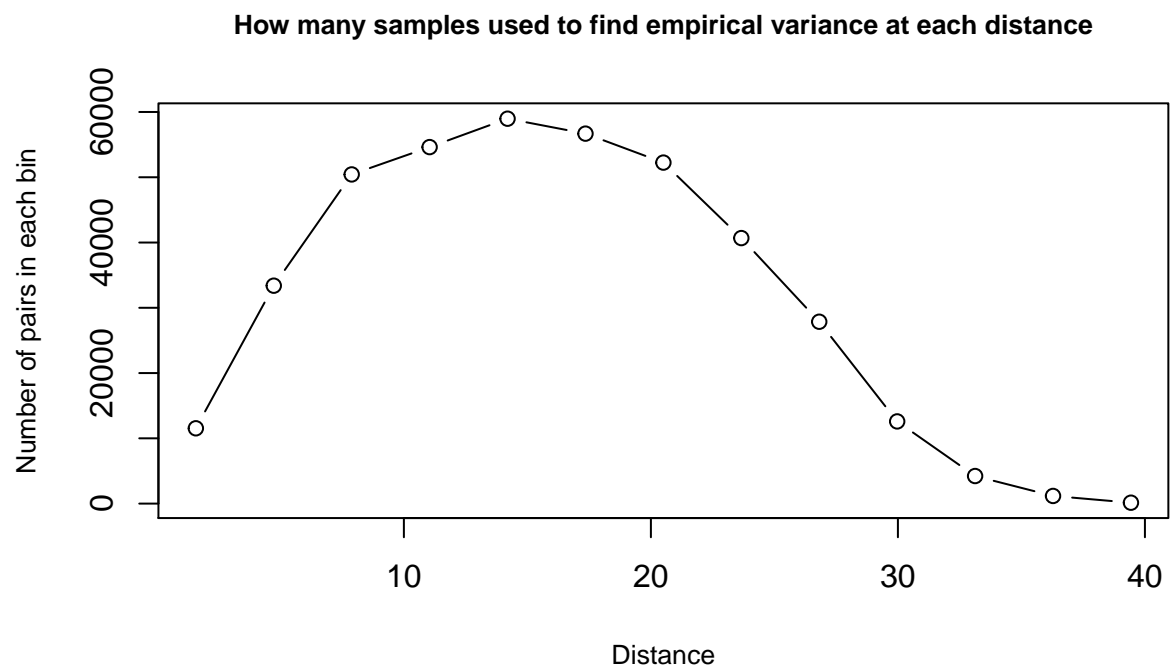
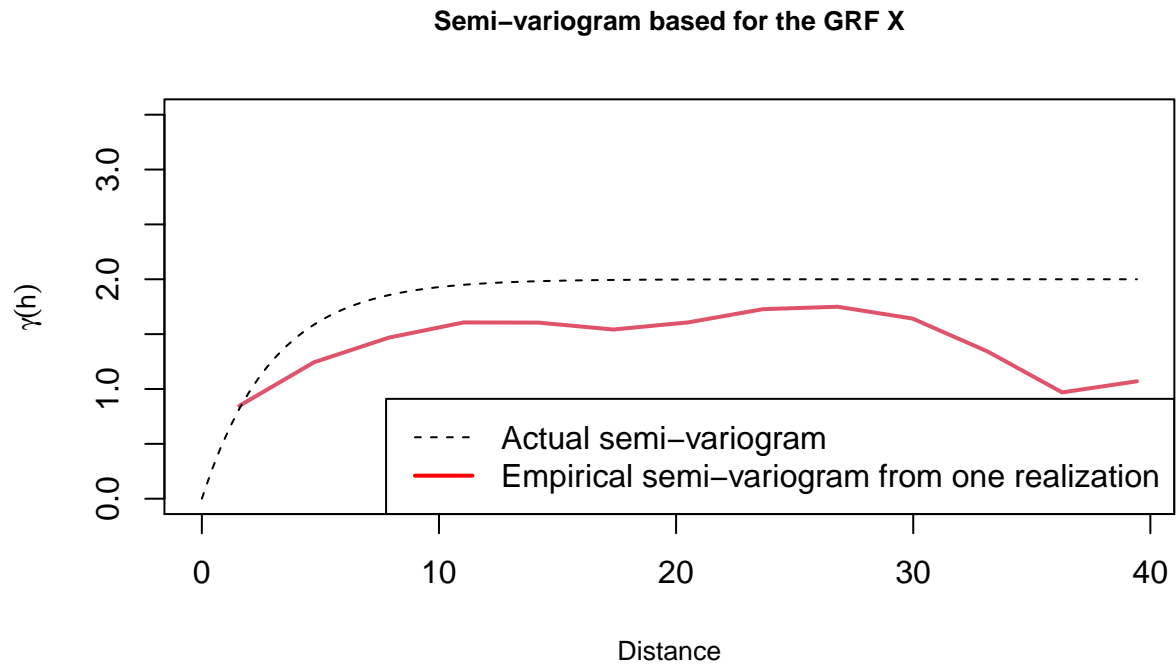
b)

We now compute the semi-variogram of our GRF, both empirical and the analytical.

```
## variog: computing omnidirectional variogram
```

```
## variog: computing omnidirectional variogram
```

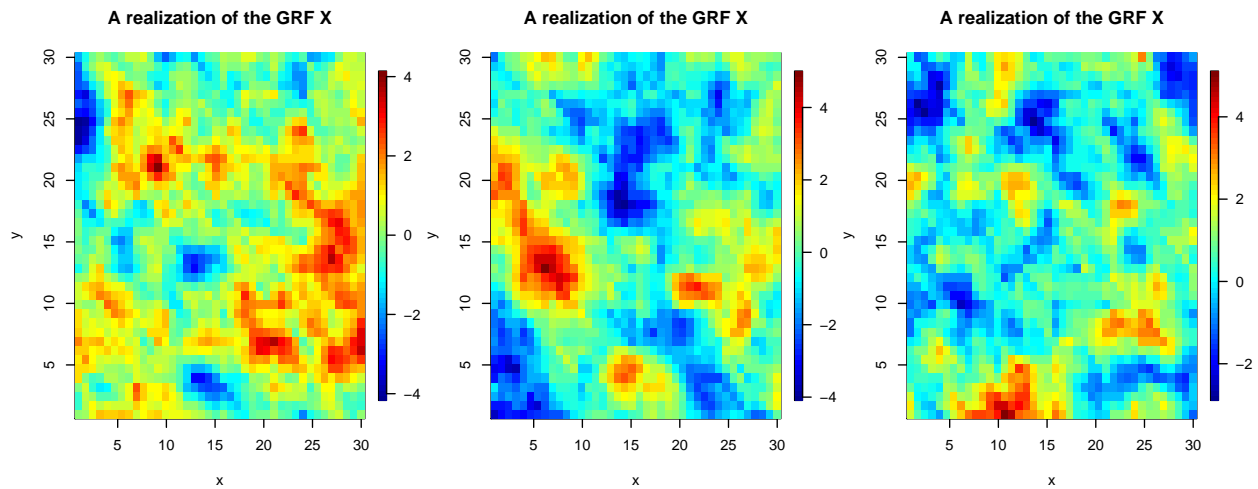
```
## variog: computing omnidirectional variogram
```



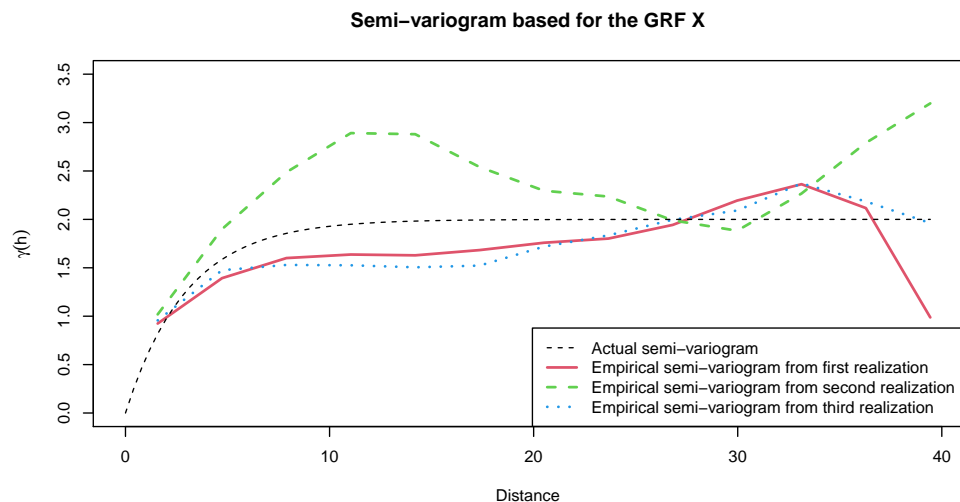
As expected, the empirical semi-variogram is more accurate for lower distances, as there are more samples to compute from. When \mathcal{D} is bounded there are fewer points that are at the distance $h \rightarrow \text{diam}(\mathcal{D})$ which in this case is approximately 40. We see also from the second plot that there is significant decrease in the amount of pairs to calculate the empirical semi-variogram from as $h \rightarrow 40$, which explains why our empirical semi-variogram is inaccurate at the end.

c)

```
## grf: simulation on a set of locations provided by the user
## grf: process with 1 covariance structure(s)
## grf: nugget effect is: tausq= 0
## grf: covariance model 1 is: powered.exponential(sigmasq=2, phi=3, kappa = 1)
## grf: simulation using the function GaussRF from package RandomFields
## ...
## grf: End of simulation procedure. Number of realizations: 3
```



```
## variog: computing omnidirectional variogram
```

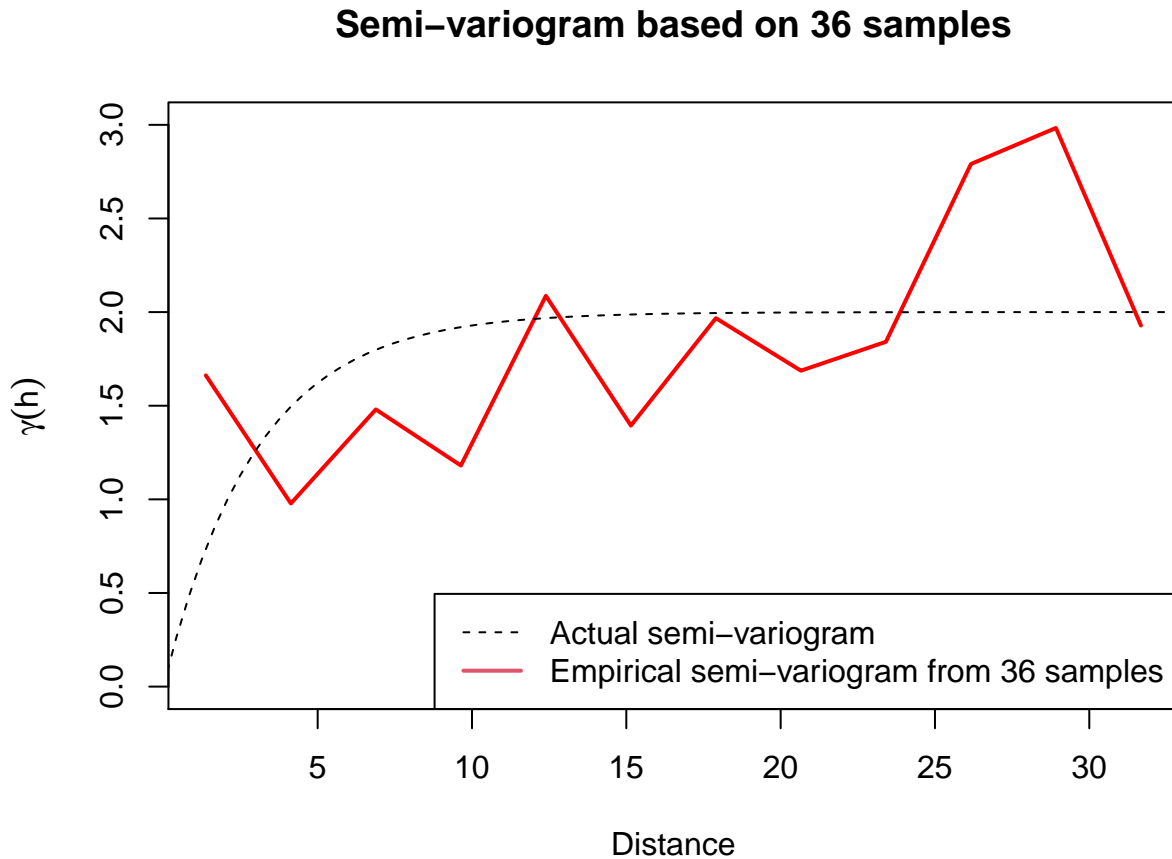


As we can see from the results, we obtain empirical semi-variograms that look similar to the actual semi-variogram. Some of them, such as the ones for the first and third realization, would seem to have a lower sill than the actual variogram, but this is just a random effect. We can see in the image plots of the GRF realizations that the variance in the GRFs seem to correspond to the values in the semi-variogram. For example, one can observe that the second empirical semi-variogram takes a large value around distance 15, and in the corresponding plot of the realization it varies more at this range than the two other plots.

The fluctuations at large distances are caused by having fewer samples at the extreme distances, as shown in the last problem.

d)

We now compare how accurate empirical estimates are for a smaller sample size. We look both at the empirical semi-variogram, as well as an estimate of the model and the maximum likelihood estimates for the covariance parameters and other eventual parameters. In our case we would expect there to be only the covariance parameters, as the exponential model doesn't take other inputs.



We first plot the semi-variogram, and then inspected this before choosing initial values in the likfit function. The sill should be the approximately the decay of the empirical semi-variogram, and we guessed the range based on where it looks like most of the dependency is gone. This was quite hard when we only had 36 samples. The true values are $\sigma^2 = 2$ and $a = 3$. Our estimate when using all the locations with initial values

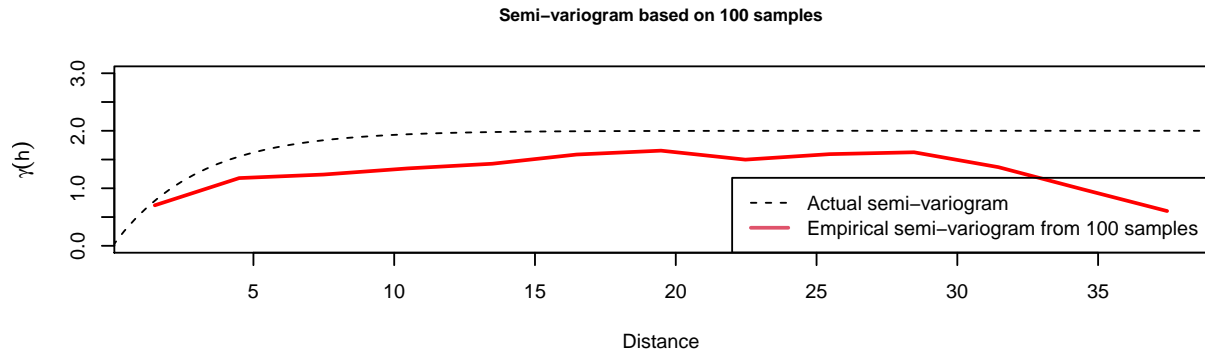
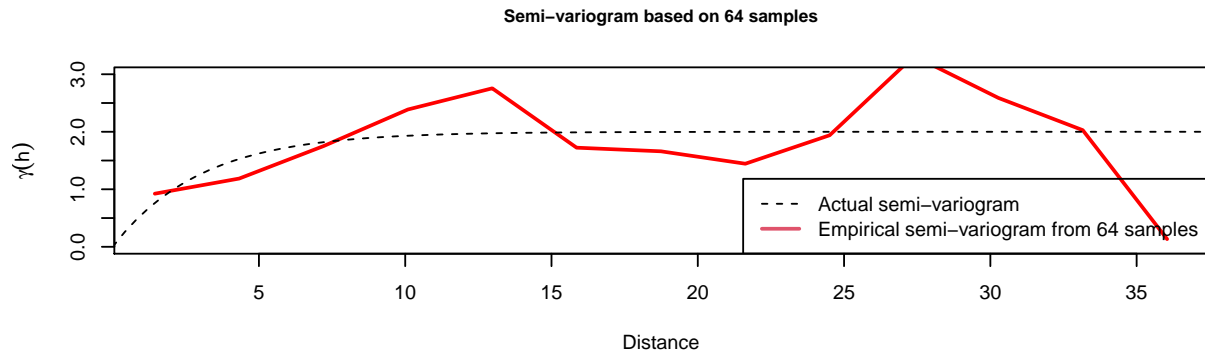
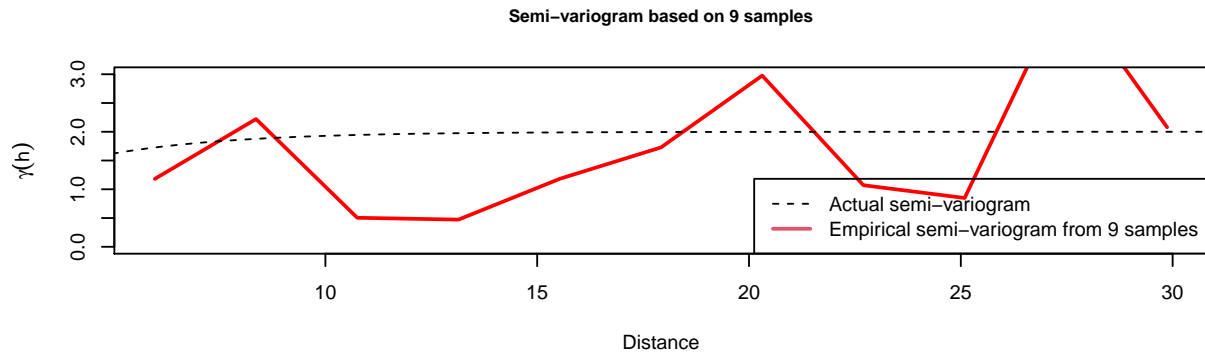
$(2, 3)$ was 1.5581282, 2.5158625, respectively, and when we used only the 36 uniformly selected locations with initial parameters $(2.5, 5)$, we obtained 0.9131814, 3.4568497, respectively.

Both likelihood based fits predicted this to be an exponential model, which is correct. If we chose worse initial parameters, the optimization would possibly run for a longer time, but it should still have a fair chance at estimating the correct parameters.

We can observe that the empirical semi-variogram we get when observing at the 36 locations becomes more spikey than the one we get when we use all locations. This is because in the latter case, we have several pairs of points with an arbitrary distance h we can use, while we risk having no locations with distance h when using only the 36 locations.

e)

We repeat the procedure with 9, 64 and 100 locations.



```
##      True values   n=9  n=36  n=64  n=100  n = all
## sigma^2          2 1.112 0.913 1.217 1.527   1.558
## a                3 0.294 3.457 3.473 3.038   2.516
```

From the table we see that these are not terrible estimates. They are a bit wrong, compared to when we use all locations, but since we guess initial parameters that are a bit off, the parameters may converge to wrong parameters. For example, the initial conditions may have been such that we end up stuck in a local maximum (for the likelihood) yielding the wrong parameters.

The semi-variogram plots are as expected, where a lower number of locations yield a smaller set of potential distances h , and hence less information about how the variance changes over distance. In conclusion the semi-variograms are not only more accurate, but also less fluctuating when we have observed the GRF at more locations.

f)

Throughout this problem, we have estimated empirical semi-variograms for different number of observations on a bounded domain. We have also estimated the (theoretically unknown) variance and spatial range of the underlying GRF.

When observing at varying number of locations, we noticed that fewer observed locations yield less pairs of locations of a given distance h , meaning we have less data to estimate our semi-variogram precisely. More locations yield more stable semi-variograms, giving more correct estimations of the sill, range and overall shape of the semi-variograms. Due to working on a bounded domain, we expect the highest distances to give a lot of variance, because in these extreme cases, there are always fewer pairs of locations with the given distance. As $h \rightarrow \text{diam}(\mathcal{D})$, the number of pairs goes to 0, yielding less and less data for estimating our semi-variogram.

Due to the semi-variograms being a little bit off compared to the true semi-variogram, we may encounter the problem that the likelihood based estimation of the initial parameters (σ^2, a) converge to the wrong value, as the optimization becomes stuck close to a local maximum. Hence, we may obtain faulty results from the likfit-estimation of the parameters, but since a higher number of locations observed often yield a more stable semi-variogram, we can often guess better initial conditions for our optimization problem when given more observations.